United States
Environmental Protection
Agency

Developing Biological Indicators: Lessons Learned from Mid-Atlantic Streams

Introduction

To demonstrate how monitoring and assessment at the regional scale could be achieved, the U.S. Environmental Protection Agency (USEPA) implemented the Mid-Atlantic Integrated Assessment (MAIA) pilot study for surface waters. A primary goal of this study was to define biological indicators that could be used to assess stream condition at the regional level. During 1993-1996, hundreds of wadeable stream sites throughout the mid-Atlantic region were surveyed for water chemistry, land use, riparian condition, and channel morphology in an effort to better understand how human influence alters fish, benthic macroinvertebrate (e.g., aquatic insect) and periphyton (e.g., algae and diatom) assemblages.

During the course of the study, researchers working independently derived different approaches to data analysis and reported different results regarding the relationships between human influence and biological change. To build consensus among the scientists involved, EPA sponsored a series of workshops to create a consistent approach for testing and selecting biological indicators for fish, macroinvertebrates and periphyton. This brochure summarizes a document that presents issues from those workshops. Please refer to *Developing Biological Indicators: Lessons Learned from Mid-Atlantic Streams* (Fore 2003, EPA/903/R-03/003) in its entirety for a more in-depth explanation of related challenges and conclusions: http://www.epa.gov/bioindicators.

To create a consistent approach for testing and selecting biological indicators

Goal of EPA-sponsored workshops

Lessons learned from the MAIA study are outlined below and highlight the steps involved in developing stream biological indicators. Efforts to resolve aspects of sampling design, data collection and management, correlating human impacts and survey data, testing and selecting individual metrics, and final development and application of a multi-metric index are presented here. This document is aimed at agency scientists or managers tasked with implementing regional monitoring programs.

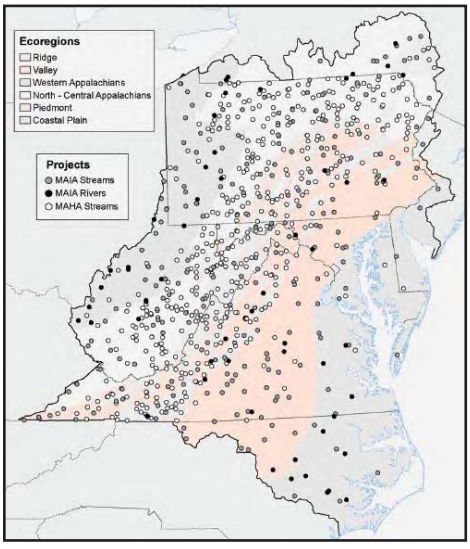


Figure 1. MAIA probabilistic study design locations for fish.

Probabilistic sampling design was the best choice for MAIA

Given the size of the sampling area and the scope of the questions asked for the MAIA study, most agreed that randomization of site selection was necessary to yield an unbiased estimate of conditions across the entire region (Figure 1). Unless one samples every site (census sampling), selecting sites randomly is the only method for inferring regional condition from a smaller set of sites.

Why sample randomly? If site selection is random and all sites are sampled with an equal or known probability, then information from the sampled sites can be used to infer the condition of sites not sampled. Thus, results based on a random sample of sites can be scaled up to the entire population of sites within a region, as long as each site in the region could have been included in the sample.

Lesson Learned:

This design strategy met the scope of the study and circumvented the need to conduct census sampling, an otherwise costly venture.

Reference sites did not always meet criteria for reference condition

Expectations for biological indicators are based on observed conditions at undisturbed or minimally disturbed locations. These reference sites are used to define reference conditions. The MAIA study used a two-pronged approach to select these sites:

- Researchers developed independent criteria, such as acid-neutralizing capacity (ANC) and nutrient loads (Table 1).
- Local biologists helped select sites based on their best professional judgment (BPJ) in order to ensure the full range of site disturbance be included, not just "typical" site disturbance.

Table 1. Independent reference site criteria.

All reference sites in calibration dataset met ALL of these criteria:

ANC > 50 µeq/L
Total Phosphorus < 20 µg/L
Total Nitrogen < 750 µg/L
Chloride < 100 µeq/L
Sulfate < 400 µeq/L
Mean RBP Score > 15

Ironically, 73% (44 out of 60) of the BPJ sites failed to meet the independently established criteria.

Lesson Learned:

Best professional judgment should always be confirmed with objective criteria in choosing reference sites.

Perils of Data Management

Different "names" for the same site caused confusion

The MAIA data set was large and complex; therefore, it was not possible to put all the data in a single file. Consistent data identifiers, particularly site names, were extremely important for matching multiple files of related data. Although the initial information management strategy accounted for this need, inconsistencies arose in the completion of data fields, which in turn complicated data analysis.

Lesson Learned:

More information should have been included within each data file to identify unique sampling occasions. Spending more time up front ensuring that data were completely and correctly stored would have saved considerable time spent trying to repair or retrieve corrupted data.

Simple files were best

Data analysis for MAIA involved multiple institutions and investigators using different statistical software. Posting files on an Internet server was the most practical approach to sharing files among so many remote users. Hosting a searchable relational database that included all the data was an option, but these were typically slow and difficult for the host to maintain. Because researchers were typically interested in a subset of data, smaller, simpler files with variables grouped according to topic worked best. The MAIA data had to be accessible to many remote users with the intention of manipulating the data within a variety of software.

Lesson Learned:

Rather than create a complicated database structure from which data would have to be exported for analysis, data files were kept simple from the beginning so that they could be easily downloaded from an EPA Internet site and quickly entered into the user's own statistical software.

Original data must be archived

The tendency was to lose track of original files with confusing formats when newer versions were created. For the MAIA study, referencing original files was the only way to catch major errors in later versions of the data.

Lesson Learned:

Original field or bench sheets must be archived along with the first generation of electronic data in a way that the data will not be changed or lost.

Linking Human Disturbance to Biological Change

Because biological systems are complex and human disturbance is multidimensional (e.g., differing types, sources, duration and intensity), single causes and mechanisms of impairment are difficult to isolate. As a result, much of the evidence for human degradation of natural resources is correlative. In such situations, although the path to causality (i.e., demonstrating cause and effect) is blocked by the inability to perform controlled experiments and use statistical inference, logical argument (or weight of evidence) constructed according to a recognized set of rules can be used instead (Table 2). In fact, this approach typically yields a stronger case because researchers consider alternative explanations explicitly, rather than ignoring them.

Results from the Mid-Atlantic illustrate how a causal argument can be constructed to support the idea that human disturbance causes biological change. Specifically, Figure 2 shows the strength of the correlation among reference site values versus sites impacted by acid deposition. This information can be used to support Beyers' first criterion for constructing causal arguments (Table 2).

Table 2. Ten criteria for constructing causal arguments.
(modified after Beyers, 1998)

1. ***Strength***: a large proportion of sampling units are affected in exposed areas compared with reference areas
2. Consistency: the association has been observed at other times and places
3. Specificity: the effect is diagnostic of exposure
4. Temporality: exposure must precede the effect in time
5. Dose-response: the intensity of the observed effect is related to the intensity of the exposure
6. Plausibility: a plausible mechanism links cause and effect
7. Evidence: a valid experiment provides strong evidence of causation
8. Analogy: similar stressors cause similar effects
9. Coherence: the causal hypothesis does not conflict with current knowledge
10. Exposure: indicators of exposure must be found in affected organisms

Beyers, D. W. 1998. Causal inference in environmental impact studies. Journal of the North American Benthological Society 17:367-173.

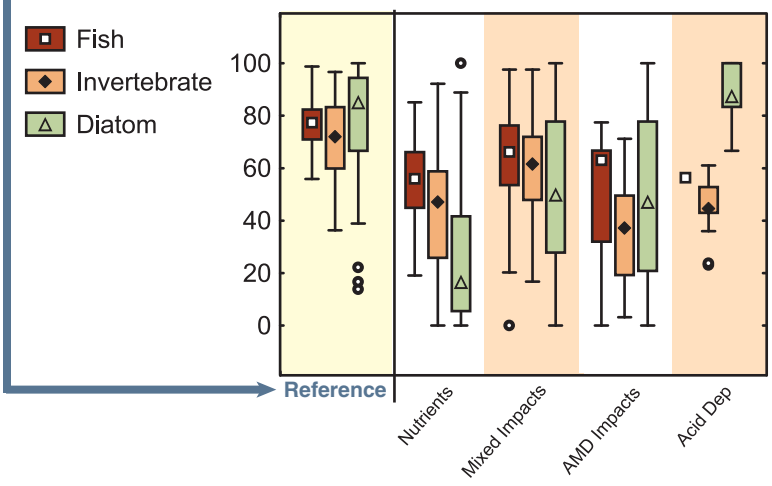


Figure 2. Multimetric index values for fish, invertebrates, and diatoms as a function of human disturbance. ***All index values were higher at reference sites***. Diatom index values were higher than invertebrate index values for sites with acid deposition.

Addressing concerns about circular reasoning

An argument based on circular reasoning is one in which the conclusion is embedded in the premise, as for example, in the statement, "decline in mayfly taxa richness is a good indicator of biological disturbance because we find many types of mayflies at undisturbed places." The concern is that the observed correlation may be due to spurious correlation with another underlying cause that drives both biology and patterns of human settlement, such as elevation or watershed size.

Lessons Learned:

In addition to the criteria in Table 2, a variety of safeguards helped reduce drawing unsubstantiated conclusions:

- Site selection was randomized across a large geographic area to ensure that the sample was representative of all possible sites.
- Measures of disturbance were selected independently of the biological metrics.
- Part of the data set was reserved to independently validate the final indexes.
- All metrics were tested for correlation with multiple gradients of human disturbance.
- Potential confounding factors such as watershed area were explicitly tested.

Patterns of human disturbance were complex

Dozens of variables related to water chemistry, metals, nutrients, fish tissue contaminants, habitat, channel morphology, geographic features, human census data, satellite land cover and use, and specific point sources were included in the data set. Hundreds more were derived from the data collected. The hope was that such a complete record of human activity would provide a clear picture of human influence and disturbance within a watershed. In reality, disturbance measures were not necessarily correlated with each other because not all activities were present in every watershed. Consequently, one of the primary challenges for the MAIA study was to determine which variables most accurately characterized human influence. Examination of a correlation matrix of all site condition variables revealed correlative patterns among related variables.

Lesson Learned:

A comprehensive study linking the types of human activities (e.g., mining or agriculture) with their specific stressors (e.g., SO_4 or nutrients) would have been helpful in clarifying metric response to disturbance. Such a study would have provided a better understanding of which measures of disturbance tended to vary together and which measures were related to natural geographic or landscape features.

Integrated versus single measures of disturbance were better predictors of human influence

Measures of generalized disturbance reflect multiple attributes of degradation as opposed to singular stressors, such as nitrogen levels or turbidity. Overall, specific stressors tended to be more highly correlated with integrative (or generalized) measures of human disturbance than they were with measures of only a single aspect of disturbance. For example, four individual measures (turbidity, pebble size, riparian vegetation condition, and riparian disturbance) were correlated with one or two of each other, but all four were correlated with Bryce, et al.'s (1999) index of disturbance, an integrated measure of site condition.

Similarly for biological indicators, integrative measures of disturbance, rather than specific stressors, showed a higher correlation with multimetric indexes for all three assemblages (Table 3). One chemical measure, chloride, was a strong indicator of general disturbance and also highly correlated with all three biological indexes.

Lesson Learned:

Measures of disturbance that integrate measures of site condition over multiple spatial scales tended to better capture the cumulative effects of human influence.

Table 3. Spearman's correlation of three multimetric indexes with selected measures of human disturbance. All correlation coefficients were significant; only values > 0.3 (or < –0.3) are shown.

Measure	Fish	Invertebrate	Diatom
N (total nitrogen)	-0.45	-0.32	-0.54
P (total phosphorus)			-0.61
NH_4 (ammonia)	-0.33	-0.36	-0.32
ANC (acid neutralizing)		-0.33	-0.53
SO_4 (sulfate)	-0.34		
Turb (turbidity)		-0.33	-0.39
%S_F (% sand and fine sediments)		-0.54	-0.39
PbSz (pebble size corrected for stream power)		0.43	0.33
RVeg (riparian vegetation)			0.30
RDist (riparian disturbance)		-0.35	
RBP (rapid bioassessment/habitat protocol)		0.42	0.36
CL (chloride)	-0.45	-0.31	-0.55
Bryce et al. [1999] disturbance categories	-0.39	-0.57	-0.54
%Dist (sum of urban, agric., & mining land use within the watershed)	-0.33	-0.40	-0.54
Total	**6**	**11**	**12**

Bryce et al. developed a risk index that summarized the intensity of human disturbance in the watershed upstream of sampled reaches. The risk index integrated information from the regional, watershed and reach scale. Each watershed was scored from 1 to 5 representing minimal to high risk of impairment.

Bryce, S. A., et al., 1999. Assessing relative risks to aquatic ecosystems: a mid-Appalachian case study. Journal of the American Water Resources Association. 35(1):23-36.

Metric Testing

Potential measures are selected for inclusion in a multimetric index if they are biologically meaningful, consistently associated with human disturbance, not redundant with other metrics, and reliably and easily quantifiable from field samples. With a large list of candidate metrics and a single test for each, there was the possibility that candidates would meet the criteria for metric selection because of chance alone. However, multiple tests against different measures of human disturbance avoided this pitfall.

Simple criteria were used first to eliminate candidate metrics

Simple statistical rules were developed to shorten the long list of candidate metrics identified for each assemblage. This first round of elimination focused on evaluating each metric's range of values, that is, the ability of a metric to differentiate levels of human disturbance.

> Lesson Learned:
>
> For Mid-Atlantic streams, candidate metrics were eliminated in favor of metrics with a broader range of values.

Statistical precision was no substitute for correlation with disturbance

Signal-to-noise ratios estimate a measure's ability to distinguish differences *among* sites from differences *within* individual sites. If the variability of a candidate metric within individual sites is higher than its variability among all sites, then the measure is unlikely to detect differences in biological condition among sites (or differences at sites that change over time).

Although most metrics incorporated into multimetric indexes have high signal-to-noise ratios (indicating high precision), a high ratio alone does not guarantee that a candidate metric will be a meaningful indicator. Metric values can be highly repeatable at individual sites but still be unrelated to human disturbance.

Consider, for example, pool depth and embeddedness. The depth of a pool in a stream is often considered an indicator of good fish habitat. Quality is expected to decline as erosion, dredging, and sedimentation fill pools, creating a homogeneous channel profile. Embeddedness represents the proportion of the stream reach filled with sand and fine sediments. For the MAIA study, pool depth measures were very precise, with signal-to-noise ratio of 16. In contrast, embeddedness measures revealed a signal-to-noise ratio of 1.9, failing to meet the authors' suggested minimum value of 2. Embeddedness, however, showed a strong correlation with human disturbance. Conversely, pool depth was precise but not related to human disturbance. Embeddedness, though less statistically precise, was the better indicator of biological condition (Figure 3).

> Lesson Learned:
>
> Certainly statistical precision is a desirable property of a good metric, but statistical precision alone does not guarantee a predictable association with human disturbance.

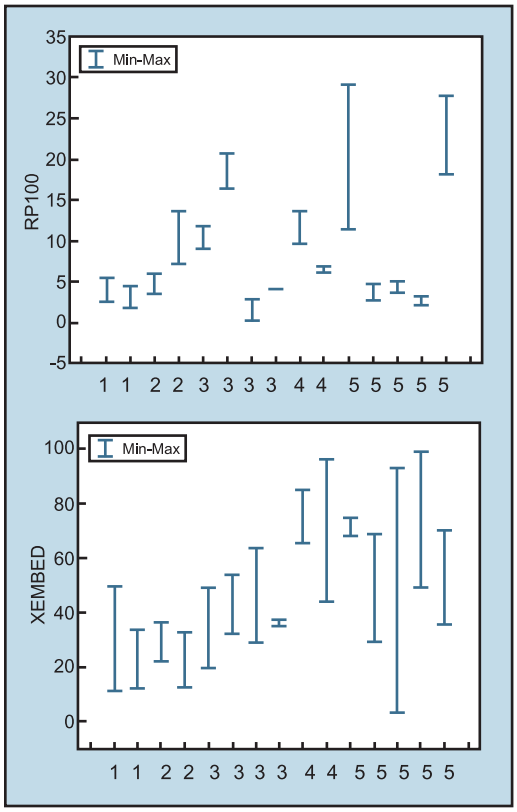


Figure 3. Ranges of values for mean residual pool depth (RP100, top panel) and embeddedness (XEMBED, bottom panel) for 15 sample sites sorted along the x-axis by disturbance class from least (1) to most (5) disturbed (Bryce et al., 1999). Vertical lines span the range of values recorded for two to six repeat visits to each site. ***Repeat visits to the same site yielded more similar values for RP100 than embeddedness indicating greater precision (shorter vertical lines); however, embeddedness consistently increased with greater human disturbance while RP100 did not.***

Metrics from different assemblages were eliminated for different reasons

The list of plausible metrics proposed for testing in Mid-Atlantic streams included 58 for fish, 120 for invertebrates, and 240 for periphyton. Most of the candidate metrics for periphyton represented untested hypotheses, whereas the other assemblage metrics had experienced a greater amount of testing. As Table 4 shows, fish metrics were eliminated for different reasons and at different frequencies than were invertebrate metrics.

Lesson Learned:

Across assemblages, metrics were selected and eliminated for different reasons.

Table 4. Numbers of candidate metrics tested for MAIA's fish and invertebrate multimetric indexes and a summary of the reasons for which they were eliminated. This **winnowing process** resulted in fewer than 10 metrics included in the final indexes.

	Fish	Invertebrates
Total # of candidate metrics	58	120
metrics eliminated: Insufficient range	13	20
metrics eliminated: Poor signal/noise	2	66
metrics eliminated: Redundant	3	25
metrics eliminated: Fail to correlate	30	2
metrics eliminated: Persistent correlation w/watershed area	1	0
Total # of metrics in final index	**9**	**7**

Development and Application of Multimetric Indexes

Multimetric indexes were created in part to fulfill a Clean Water Act mandate that all states develop numeric criteria for assessing biological condition of water bodies. A multimetric index, as the name implies, is a carefully constructed framework of multiple types of measurements. Once individual metrics have been tested and selected for inclusion in a multimetric index, it is necessary to ensure the index as a whole will offer a reliable and quantifiable indication of human disturbance.

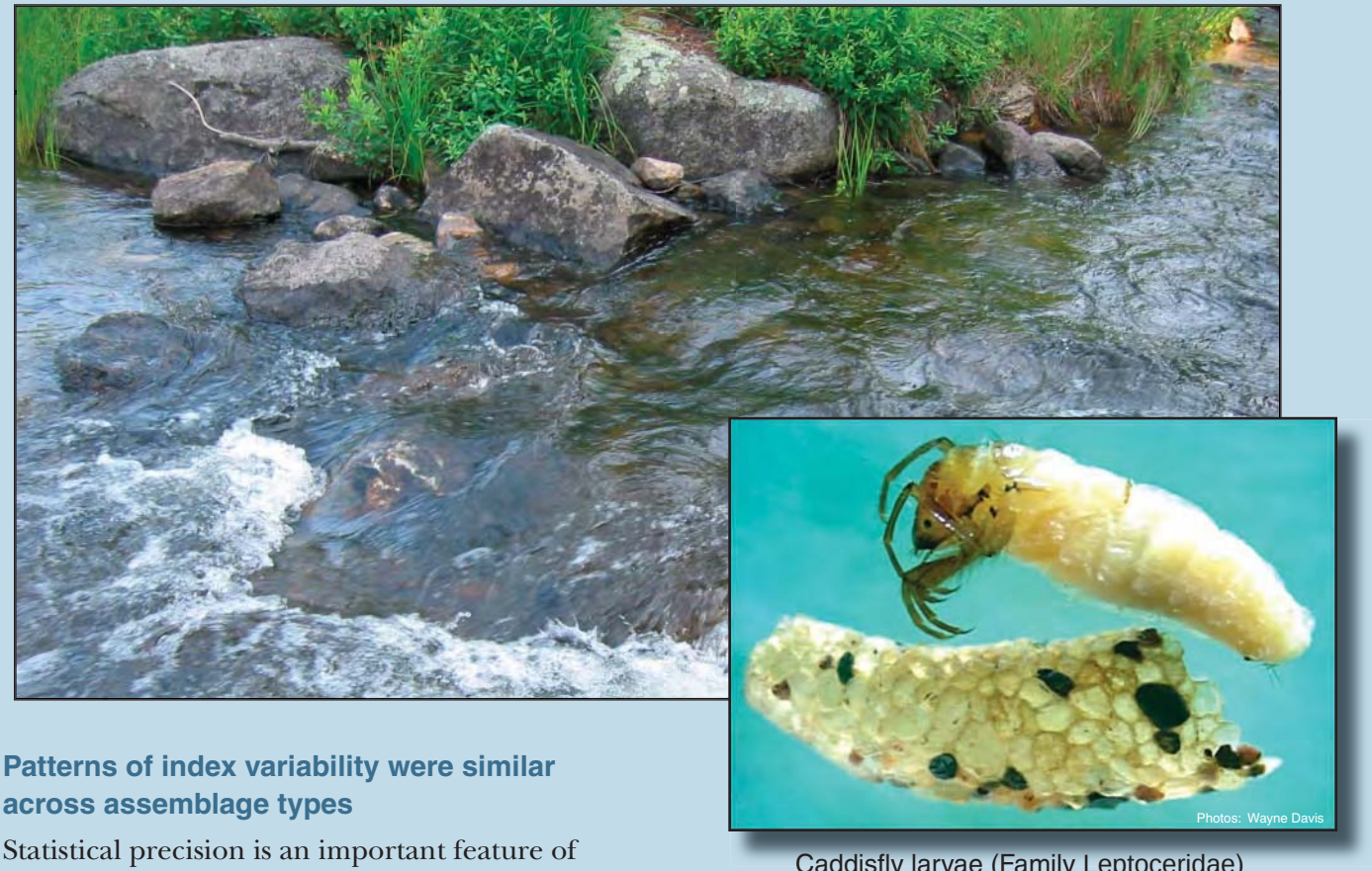
Biological criteria depend on the definition of reference sites

In the same way reference sites are used to develop individual metrics, multimetric index values observed at reference, or minimally disturbed, sites are used by many states to define biological impairment. Currently, states vary both in the way they characterize reference condition and define deviation from reference condition. States also vary in their determinations of biological impairment thresholds.

Lesson Learned:

From the MAIA study we learned the importance of having objective criteria to select reference sites, a lesson relevant to all states as they develop reference condition criteria and rules for defining impairment. Additionally, it is important to develop informative, defensible and consistent thresholds from state to state.

Red breasted sunfish



Caddisfly larvae (Family Leptoceridae)

Patterns of index variability were similar across assemblage types

Statistical precision is an important feature of any monitoring tool because it determines the ability of an indicator to detect change should it occur. A highly variable indicator must show a large change in value before the change is statistically significant. Lack of sensitivity translates into an inability to sound an alarm that will protect resources from degradation.

Statistical power analysis can be used to estimate the magnitude of change that an indicator can detect. Results from two commonly used statistical models for power analysis (t-test and regression) indicated that the MAIA multimetric indexes had adequate precision to distinguish between two and five categories of biological condition (such as good, fair, poor) and could detect between 1.5% and 2.5% change per year after five years of monitoring.

As shown in Table 5, indexes for each assemblage differed in percentages of "nuisance" variance, that is, the amount of an index's total variance that can be explained by year-to-year differences, statistical outliers, and measurement error. Site variance is associated with biological condition—the higher the percentage of site variance, the more precise the index.

Lesson Learned:

Despite differences in how the statistical models ranked the three indexes, percentages of "nuisance" variance components were approximately similar across assemblages (13-20%).

Table 5. Components of variance expressed as a percentage of the total variance for diatom, invertebrate, and fish multimetric indexes. Variance associated with site differences, year-to-year differences, site x year interaction, and repeat visits within years are shown for each index.

		Percentage of total variance		
		Diatom	Invertebrates	Fish
	Site (target variability)	80.4	83.3	86.8
"nuisance"	Year (year-to-year differences)	0	2.1	1.5
"nuisance"	Site x year (statistical outliers)	2.7	1.6	5.6
"nuisance"	Error [repeat visits] (measurement error)	16.9	12.9	6.2

U.S. Environmental Protection Agency
Mid-Atlantic Integrated Assessment
701 Mapes Road
Fort Meade, MD 20755-5350

EPA/903/F-06/001
January 2006

Assemblages differed in their sensitivity to disturbance types

The MAIA study concluded that any of the three assemblages could be used to monitor stream condition because multimetric indexes for all three assemblages could reliably distinguish degraded sites from sites with little or no human influence. However, each assemblage varied in its sensitivities to different types of disturbance. Figure 4 shows the relative sensitivity to disturbance conditions (or relative risk) of each assemblage. For example, fish showed less sensitivity to sedimentation effects than invertebrates or algae.

Lesson Learned:

Employing all three multimetric indexes to monitor stream condition yields the fullest range of information.

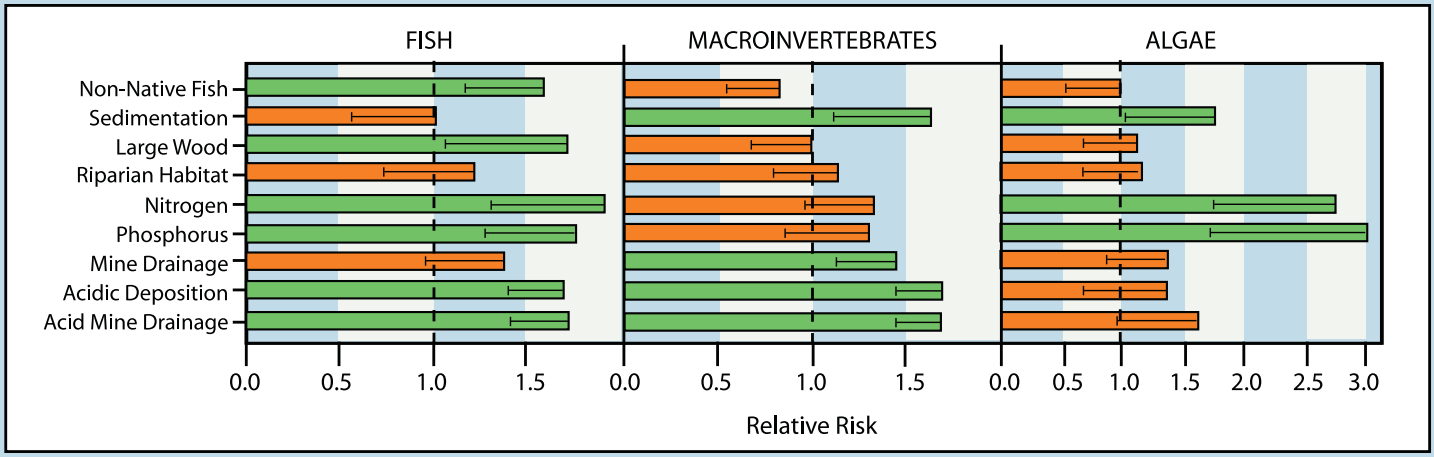


Figure developed by John Stoddard based on data from the Mid-Atlantic Highlands Streams Assessment (EPA 2000, EPA/903/R-00/015)

Figure 4. A relative risk of 1.0 denotes "no stressor effect", and ***stressors with confidence intervals lying entirely above 1.0 (green bars) are statistically significant (one-sided p<=0.05).*** This figure shows relative risk values for associations between biotic integrity (for each assemblage) and stressor condition (for each assessed stressor). Length of bars is the increase in likelihood of encountering a poor ecological condition (based on biological indicators) when the stressor is also ranked as poor.

Contact

Wayne Davis
Office of Environmental Information
Environmental Analysis Division
Mid-Atlantic Integrated Assessment
701 Mapes Road
Fort Meade, MD 20755-5350
410-305-3030
davis.wayne@epa.gov
www.epa.gov/bioindicators

Lou Reynolds
USEPA Region 3
EAID Freshwater Biology Team
1060 Chapline Street
Suite 303
Wheeling, WV 26003-2995
304-234-0244
reynolds.louis@epa.gov

John Stoddard
USEPA Office of Research and Development
NHEERL - Western Ecology Division
200 SW 35th Street
Corvallis, OR 97333
541-754-4441
stoddard.john@epa.gov
www.epa.gov/nheerl/arm